跨情境的刺激泛化在面孔信任形成中的作用：基于直接互动与观察学习的视角*

袁博 王晓萍 尹军 李伟强

（宁波大学心理学系暨研究所, 浙江 宁波 315211）

摘 要 基于联结学习理论(associative learning theory)，通过 3 项实验考察了跨情境（公平—信任）的刺激泛化在面孔信任形成中的作用。实验 1a 和实验 1b 分别从直接互动和观察学习视角，发现了面孔信任形成中跨情境的刺激泛化效应(stimulus generalization effect)，即相比于中等不公平条件，随着被信任者的面孔与先前互动中公平（不公平）分配者面孔相似度的增加，个体对其信任程度逐渐增加（降低）；并且这一效应具有不对称性(asymmetry)，对不公平分配者面孔的泛化强度高于对公平分配者面孔的泛化强度。采用漂移扩散模型(Drift-Diffusion Modeling, DDM)分析发现，不公平条件下的漂移率v显著小于中等不公平或公平条件下的漂移率v，且大多分布在小于 0 区间；表明在对与先前互动中不公平分配者面孔相似的陌生面孔进行信任决策时，个体更倾向累积不信任的证据。实验 2 结果发现，行为意图在刺激泛化效应的产生中起到调节作用；在无意图条件下，上述跨情境的刺激泛化效应消失。上述结果表明，个体采用联结学习机制将不同情境中习得的刺激价值联结泛化到新的互动情境中，进而指导随后的信任决策。

关键词 信任形成，联结学习，刺激泛化，行为意图，漂移扩散模型

1 引言

人类在社会环境中的生存和发展离不开社会交互(高青林, 周媛, 2021; Radell et al., 2016)，而社会交互的发起和维系需要进行适应性的信任决策，即评估他人是否值得信赖，进而决定是否信任他人(Ruff & Fehr, 2014)。信任(trust)是建立在对他人的意向或行为的积极预期基础上，敢于托付(愿意承受风险)的一种意愿，也是人类社会交互的基本组成部分(Rousseau et al., 1998)。人们可以根据与他人的直接交互经验来决定是否信任，例如，多轮信任游戏(multi-round trust game)的研究表明，经历积极的信任体验会增加个体对他人的信任，而经历信任违背则会降低个体对他人的信任(Erev & Roth, 1998)。但当遇到不熟悉的他人时，即使缺乏与

收稿日期: 2021-12-30

* 全国教育科学规划一般项目“青少年道德决策中的同伴影响及其认知情感机制研究”(BBA210033)

通讯作者：袁博, E-mail: yuanbopsy@gmail.com; 李伟强, E-mail: liweiqiang@nbu.edu.cn

此人的直接互动经验，人们依然会迅速做出是否信任对方的决定，并承担由此带来的经济和情感后果。在这种缺乏直接互动经验的情况下，是什么决定着我们的信任决策?

面孔作为与他人互动时的首要线索，在人际交往中发挥着重要作用。在缺乏直接经验或声誉(reputation)信息的情况下，面孔本身的一些特征会对个体的信任决策产生影响。例如，Willis 和 Todorov(2006)研究发现，人们能够在 100ms 内快速判断面孔的可信度信息，即使提供更多加工时间，这种判断也基本保持不变。已有研究发现，眼睛、眉毛、下颌等部位的特征会显著影响个体对面孔的信任决策行为。比如，具有圆脸、大眼睛的个体更可能被知觉和判断为可信的人，而那些具有颧骨下陷等特点的人更可能被知觉和判断为不可信的人(Todorov et al., 2008)。也有研究发现，面孔吸引力(attraction)会对个体的信任决策行为产生影响。与面孔吸引力较低的个体相比，面孔吸引力较高的个体在信任游戏中获得“美丽溢价”，被信任的比例更高(Wilson & Eckel, 2006)。这些研究表明，人们会依据面孔本身的一些特征，对其做出信任决策。

当面孔本身包含的信任特征有限时，个体会依赖哪些其他信息对陌生他人做出信任决策?联结学习理论(associative learning theory)认为，先前的经验会用于指导个体在随后互动中进行基于价值(value based)的决策选择(Rescorla & Solomon, 1967; Dayan & Berridge, 2014)。其中，刺激泛化(stimulus generalization)是指价值可以在感知上或概念上彼此相似的刺激之间传播和转移(Rescorla, 1976; Verosky & Todorov, 2010)。产生刺激泛化的基础是两个刺激在某一属性上的相似性，比如，当陌生面孔与熟悉面孔共享某些属性时，熟悉面孔的的心理表征被自发激活，从而导致个体对陌生面孔产生与熟悉面孔相同的评价(Kraus et al., 2010)。已有研究发现，刺激泛化在对陌生面孔的印象形成中发挥了重要作用(Kocsor & Bereczkei, 2017)。如，Gawronski 和 Quinn(2013)研究发现，当熟悉面孔具有积极/消极特质时，个体对与其相似的陌生面孔的积极/消极评价会随相似度的增加而增加。因此，当陌生面孔本身的特征无法为个体提供信息时，陌生面孔与熟悉面孔的相似性就成为个体对陌生面孔印象形成的重要线索。那么，在信任决策中，是否也存在刺激的泛化效应，即未知面孔和已知面孔的相似性是否以及如何影响人们的信任决策？

Feldmanhall 等人(2018)采用改编的重复信任博弈(repeated trust game)范式，考察了刺激泛化在信任形成中的作用。在刺激联结阶段(conditioning phases)，被试作为投资者对 3 名预先评定的具有中等吸引力、可信度、领导力的三张面孔进行投资。投资的钱将会翻三倍至被投资者，随后被投资者可以选择保留全部收益，也可以选择将收益的一半返还给投资者。在实验中，可信的被投资者有 93%的可能性选择返还收益；中等可信的被投资者有 60%的可能性选择返还收益；不可信的被投资者仅有 7%的可能性选择返还收益。在刺激泛化阶段(generalization phases)，被试可以选择一张未知面孔或与刺激联结阶段的 3 名被投资者具有不同相似度的变形(morph)面孔进行接下来的信任游戏。结果发现，个体会选取与刺激联结阶段值得信赖的面孔相似的面孔进行信任游戏，而避免与刺激联结阶段不值得信赖的面孔相

似的面孔进行信任游戏。上述结果表明，信任决策中也存在刺激泛化效应，人们利用刺激泛化机制指导其做出适应性的信任决策。

然而，Feldmanhall 等人(2018)的研究仅探讨了同一情境下（刺激联结和泛化阶段均为信任博弈任务），通过直接互动产生的刺激泛化在信任形成中的作用。在现实生活中，社会互动情境复杂多变，刺激很少在完全相同的情境中出现，先前形成刺激联结的情境通常与随后产生刺激泛化的情境差异很大。例如，人们可能在先前的互动情境中与他人进行资源分配，在随后的情境中做出信任决策。人们能否从不同情境中学习互动对象的声誉信息，并将其泛化到随后对陌生面孔的信任决策？除了直接互动学习，人们能否通过观察学习形成类似的刺激泛化用于指导随后的信任决策？此外，以往研究表明，行为意图(intention)在道德判断或社会互动中起到重要作用(Cushman et al., 2009; Ma et al., 2015)。那么，面孔信任形成中的刺激泛化仅需要刺激与行为结果之间的简单联结，还是需要基于对他人行为意图的感知？对于上述问题进行探讨，将有助于拓展以往联结学习的研究，揭示社会互动中面孔信任形成的心理机制。为此，**本研究拟从直接互动和观察学习视角，探讨跨情境间的刺激泛化在面孔信任形成中的作用，以及行为意图在刺激泛化中的作用。**

学习他人声誉信息的能力需要个体对先前互动中形成的道德印象高度敏感。公平(fairness)作为人类社会最核心的一则社会规范，对个人的生存和社会的稳定都至关重要。人们对社会互动中的公平性非常敏感，具有强烈的不公平厌恶倾向(inequity aversion)；当受到不公平对待时，会产生不满和怨恨，并不惜牺牲自己的经济利益以惩罚对方(Chang & Sanfey, 2013; Xiang et al., 2013)。研究发现，社会互动的不公平经历会减弱个体对不公平实施者的共情(Singer et al., 2006)；相对游戏中表现不公平的搭档，个体更愿意信任那些游戏中表现公平的搭档(Fareri et al., 2012)。除了直接互动学习，个体还通过观察他人行动的结果，间接地习得关于刺激的价值，并调整自己未来面对相同刺激时的行为反应，即观察学习(observing learning) (Shang & Li, 2020; Olsson et al., 2020)。日常生活中，我们不仅会亲身经历不公平的事件，还会作为第三方观察到不公平事件。采用第三方惩罚(third-party punishment)范式的研究表明，当个体观察到不公平行为时，也会产生不公平的情绪体验，并且可能损失自己的利益惩罚不公平的分配者(Buckholtz et al., 2008; Mcauliffe et al., 2015)。作为社会性动物，人类会通过观察他人在社会互动中的行为表现，获取他人在社会层面的效价信息(Earley, 2010; 郑旭涛　等, 2020)，形成对他人的声誉表征，并在随后的互动中根据他人声誉做出相应行为调整(Milinski, 2016)。鉴于公平在社会互动中的重要作用，个体对互动对象表现出的公平性非常敏感。**我们推测，个体会通过直接互动或观察学习，对社会互动中表现出不同公平程度的互动对象的面孔形成相应的价值联结，并对他们的面孔产生刺激泛化，用以指导在随后的互动中对陌生面孔的信任决策。**

以往研究发现，在令人厌恶的领域(aversive domains)存在更强的泛化，因为通常错误地将危险刺激识别为安全的比将安全刺激视为危险的代价更大(Bateson et al., 2011)。在对社会

行为的观察学习中也普遍存在消极偏向(negativity bias)，即相比于积极刺激，人们更容易注意消极刺激，进而形成更加复杂的认知表征(Baumeister et al., 2001; Rozin & Royzman, 2001)。Schechtman 等人(2010)研究发现，与中性和积极的声音信息相比，消极的声音信息引发更强的泛化。研究者认为，相比于消极事件，积极事件发生概率更高，效价评估基线标准受其影响向积极方向偏移，导致消极事件的吸引力增强，且消极事件对人类威胁更大，对其感知更敏感(Baumeister et al., 2001; Rozin & Royzman, 2001)。Feldmanhall 等人(2018)的研究也发现，与值得信任的面孔相比，被试对不值得信任的面孔存在着更强的泛化，表现出刺激泛化的不对称性(asymmetry)。**因此，我们进一步假设，在直接互动或间接观察学习下，对不同公平程度的面孔泛化中也存在不对称性，个体对不公平面孔的泛化程度要强于对公平面孔的泛化程度。**

泛化发生的前提是刺激与价值或情感上的紧密联系(王天鸿 等, 2020)，哪些因素可能促进或抑制泛化效应的发生？大量研究发现，道德印象的形成不仅受到行为结果的影响，还会受到行为意图的调节(Offerman, 2002; Dufwenberg & Kirchsteiger, 2004)。Cushman 等人(2009)研究发现，随着年龄的增长，人们会更多的基于意图对道德行为进行判断；相比于无意图，恶意意图通常被赋予更多的不道德效价。Sutter(2007)的研究发现，与不公平的结果相比，不公平的意图对个体行为的影响更大，人们更有可能拒绝具有不公平意图的分配方案。也有研究发现，在最后通牒博弈(ultimatum game)中，对于人和计算机给出的相同数额的分配方案，回应者会更频繁的拒绝人提出的不公平方案(Moretti & Di Pellegrino, 2010; Van't Wout et al., 2006)。Sun 等人(2020)采用改编的两轮独裁者博弈(dictator game)范式，探讨了分配者的行为意图对接受者间接互惠行为的影响。结果发现，仅在意图条件下，先前的公平或不公平经历才会影响被试随后的间接互惠行为。上述结果表明，行为意图会调节个体对分配结果的公平性判断及其随后的行为反应。据此，**我们推测，行为意图在刺激泛化效应的产生中起到调节作用，仅在有意图的条件下，对先前互动中的面孔才会形成相应的刺激价值联结，并泛化到随后的面孔信任决策中。**

综上，本研究拟通过 3 项实验考察跨情境下（公平−信任）的刺激泛化在面孔信任形成中的作用。实验 1a 和实验 1b 分别从直接互动和观察学习视角，探讨与不同公平程度分配者的面孔形成的价值联结，是否会影响个体对与之相似陌生面孔的信任决策。实验 2 进一步探讨行为意图是否调节刺激泛化效应的产生。为此，我们在实验中设置了刺激联结和刺激泛化两个阶段。在刺激联结阶段，被试与 3 名不同公平程度（公平、中等不公平、不公平）的分配者进行最后通牒博弈，让被试形成对 3 名分配者面孔的刺激价值联结。在随后的刺激泛化阶段，被试在两个面孔（变形 vs.匿名）之间选择其中一个作为信任游戏的搭档。通过操纵变形面孔与之前最后通牒博弈中分配者面孔的相似度，检测先前互动中与不同公平程度分配者面孔形成的刺激价值联结，是否会泛化到对不同任务情境下知觉相似的变形面孔的信任决策中。

2 实验1a 直接互动下，跨情境的刺激泛化在面孔信任形成中的作用

2.1 被试

基于本研究的实验设计，设定显著性水平 $\alpha = 0.05$，统计检验力 $1\text{-}\beta = 0.80$，达到中等左右效应量 $f = 0.25$，根据 G*power 计算，每个条件下需 28 名被试。因此，在某高校的学生群体中，招募 31 名在校大学生，被试平均年龄为 21.41 ($SD = 1.60$)。其中男生 16 人，女生 15 人。被试实验前签署知情同意书，实验后给予一定的实验报酬。

2.2 变形(morph)面孔刺激的构建

首先，为了选取具有中等吸引力、可信度、领导力的面孔，随机选取某高校 32 名被试对预先从 CAS-PEAL 人脸数据库中筛选出的 20 张面孔进行评定(Gao et al., 2007)。最终选择出在吸引力、可信度和领导力三个维度相匹配的 3 张原始面孔图片，以及另外 6 张具有中等可信度和中等吸引力的面孔图片。

随后，使用 Abrosoft Fanta Morph 软件(https://www.fantamorph.com/index.html)对上述面孔图片进行变形，将在吸引力、可信度和领导力三个维度相匹配的 3 张原始面孔，以 11%为增量（12%、23%、34%、45%、56%、67%、78%、89%），分别与预先评定为中等可信度和吸引力的 6 张面孔进行变形。使得原始的 3 张面孔能够分别产生 48 个新的面孔刺激，共产生 144 个新面孔图片。为了避免被试觉察到面孔是通过变形而产生的，我们删除了与原始面孔过于相似（89%）和差异过大（12%）的变形刺激，剩下变形增量为 23%、34%、45%、56%、67%、78%的 6 种变形面孔。这样，原始的 3 张面孔能够分别产生 36 个新面孔刺激，共 108 张面孔刺激。

此外，为避免被试发现面孔之间过于相似，我们只使用了沿同一连续体彼此相距两个增量的变形面孔，即同一连续体上两个相邻的变形面孔之间的变形差异为 22%（如，选择 23%、45%、67%或 34%、56%、78%的变形增量，见图 1）。最终，每种原始面孔对应有 18 张新面孔，共有 54 张。



图 1 面孔刺激构建过程示意图。(A) 原始面孔与 6 张预先评定的面孔以 11%为增量分别进行 morph 变形。为避免被试发现面孔之间过于相似，只选择了沿同一连续体彼此相距两个增量的变形面孔。(B) 最终用于实验的一张原始面孔以及与之相对应的变形面孔。

2.3 实验流程

2.3.1 刺激联结阶段

刺激联结阶段采用最后通牒博弈任务，让被试对 3 名不同公平程度的分配者面孔形成价值联结。分配者的面孔图片来自于预先评定的 3 张原始面孔，面孔和效价的联结在被试间平衡。3 名分配者分别对应着 3 种不同公平程度的分配条件：公平的分配者会提出 6 次 10/10、2 次 9/11、2 次 8/12 的分配方案；中等不公平的分配者会提出 6 次 7/13、2 次 6/14、2 次 5/15 的分配方案；不公平的分配者会提出 6 次 1/19、2 次 2/18、2 次 3/17 的分配方案。具体流程如下：首先，屏幕上会呈现一个“+”，呈现时间为 600ms，提示本轮实验任务开始。随后，屏幕上会呈现分配者的面孔图片，被试有 2000ms 的时间进行观察。接着 600ms 的空屏之后，呈现分配者面孔及其提出的分配方案。被试需要决定是否接受该分配方案，如果接受按“*F*”键，拒绝按“*J*”键。做出按键反应后，会对被试的选择结果进行反馈，反馈呈现 2000ms（如图 2 所示）。此外，在 30%的试次中，被试还需要对看到该分配方案后的愉悦情绪进行 1~9 点(*1 = 非常不高兴, 9 = 非常高兴*)评定（实验 1a 具体指导语见补充材料）。

为了检验刺激联结的效果，确定被试能否将 3 名分配者的面孔与不同公平程度之间形成联结。在所有试次结束后，分别随机呈现 3 名分配者的面孔，被试需要对每名分配者在先前任务中表现出的公平程度进行 1~9 点(*1 = 非常不公平, 9 = 非常公平*)评定。完成该阶段任务大约需 10 分钟。

2.3.2 刺激泛化阶段

刺激泛化阶段旨在考察与先前互动中不同效价面孔形成的联结，是否会泛化到不同情境下知觉相似的面孔中。在该阶段任务中，被试（投资者）从一张匿名面孔和一张变形面孔中选择信任投资的搭档（被投资者）。其中，变形面孔来自于变形面孔刺激的构建部分所描述的 54 张面孔，这些面孔与原始面孔的相似度分别为：23%、34%、45%、56%、67%、78%。匿名面孔是一张人脸的灰色轮廓图，表示计算机将随机匹配给被试一名“被投资者”。具体流程如下：首先，屏幕上会呈现一个“+”，呈现时间为 800ms，提示本轮实验任务开始。随后，呈现一张变形面孔和一张匿名面孔图片，变形面孔和匿名面孔在左右两侧进行平衡。被试决定选择哪张面孔图片进行接下来的信任游戏，按“*F*”键表示选择左边的面孔，按“*J*”键表示选择右边的面孔进行接下来的信任游戏。被试做出按键选择后，对被试的选择加以红框确认，呈现 2000ms（见图 2）。被试对任务理解无误后，开始实验正式，完成该阶段任务大约需 15 分钟。整个实验程序采用 PsychoPy 软件(Peirce, 2009)编写。



图 2 实验 1a 中刺激联结和刺激泛化阶段的流程图

2.4 数据分析处理

为检验与不同效价面孔形成联结后，面孔相似性是否会影响被试对陌生面孔的信任决策行为。将面孔联结的类型（公平、中等不公平、不公平）、面孔相似度（23%、34%、45%、56%、67%、78%）作为预测变量，以被试选择变形面孔的比率为响应变量进行混合线性回归(mixed linear regression)分析。混合线性模型包含了随机效应，提高了统计检验力，并且能够处理违反方差齐性假设的数据(Baayen et al., 2008)。参照 Feldmanhall 等人(2018)的分析，对面孔联结类型进行了哑变量处理，将中等不公平作为参照水平，以便考察不公平和公平条件下的刺激泛化效应。使用基于 R 语言的 lme4 包(Boeck et al., 2011)进行分析，将被试作为随机效应，面孔联结的类型和面孔相似度作为固定效应。此外，我们采用 brms 包(Bürkner, 2017)进行了贝叶斯线性混合模型分析，固定效应与随机效应处理与上述混合线性回归分析相同。使用 brms 包默认的先验分布，模型拟合均使用 4 条独立的 MCMC 链，每条链包含 2000 个有效样本。贝叶斯回归分析可以直接计算出回归系数后验分布(posterior distribution)的可信度区间(credible intervals, *CIs*)，进而进行统计推断，比如，95% *CIs* 是否包括 0。

对于反应时数据，我们首先进行了传统的反应时数据分析，其次采用漂移扩散模型(Drift-Diffusion Modeling, DDM)对反应时数据进行了分析。DDM 把决策描述为一个连续的抽样过程，即带有噪声的信息从起点累积到对应于某一选项的边界或阈值，该选项被选中(Ratcliff & McKoon, 2008)。DDM 模型参数包括漂移率(drift rate, v)、起始点偏差(bias, z)、边界高度(boundary, α)和非决策时间(non-decision-time, τ)。其中，漂移率v代表累积某一选择证据的速率，即个体倾向于某一选项的偏好越强烈，信息向该选项积累的速度就越快；起始点偏差z表示决策之前的先验偏向；边界高度α表示在做出反应之前必须积累的信息量；非决策时间τ反映了影响决策反应时中的其他因素，包括信息编码与按键反应的时间(张银花等, 2020)。DDM 可以将潜在的认知过程体现在模型不同的成分上，从而将决策过程分解为具有心理意义的参数(Johnson et al., 2017)。比如，在本研究中，DDM 使用选择和反应时分布来描述被试如何累积证据做出信任选择。漂移率v量化了被试通过加工面孔信息获得的有利于选择信任或不信任证据的强度，即对选择信任或不信任的价值权衡程度[1]。起始点偏差z量化了被试在获取任何证据之前选择信任/不信任的倾向。边界高度α量化了被试在做出选择时需要的证据量，进而反映了不同条件下个体做出选择的谨慎程度（见图 3）。

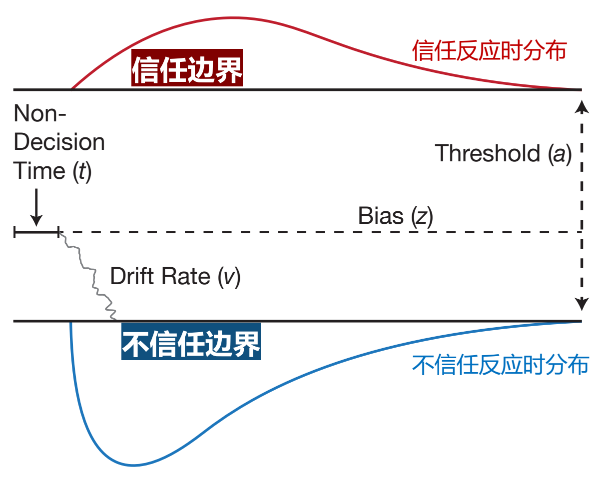


图 3 个体做出信任选择时，漂移扩散模型的示意图

采用 Python 软件包 HDDM(hierarchical drift diffusion model)对反应时数据进行 DDM 模型分析(Wiecki et al., 2013)。HDDM 使用分层贝叶斯参数估计(hierarchical Bayesian parameter estimation)的方法同时拟合个体(individual)与群体(group)层面的参数。因此，可以直接对估计出的后验(posterior)参数进行差异性检验，而不用依赖于传统的频率统计分析。比如，可以选取两个感兴趣条件下参数的后验分布，如不公平面孔联结和公平面孔联结条件下的漂移率；然后计算每个抽样中两个条件的参数值之间的差异，产生一个可信的平均差异分布(credible mean differences)。如果这个分布的 95%的最高密度区间(highest density interval, HDI)不包含 0，那么两个条件之间的差异就是可信的(Johnson et al., 2017)。其余实验的 DDM 数据分析方

[1] 在本研究的 DDM 中，我们将选择信任编码为 1，选择不信任编码为 0。

法基本与实验 1a 一致，下文不再赘述。

2.5 结果

2.5.1 刺激联结阶段

为了检验刺激联结的效果，对被试在不同条件下的接受率、分配方案引发的愉悦情绪、以及对分配者面孔的公平程度评分进行单因素被试内方差分析。结果发现，不同条件下对分配者提议的接受率差异显著，$F(2, 60) = 135.56$，$p < 0.001$, $\eta^2 = 0.82$。对公平分配者提议的接受率($M = 0.95$, $SD = 0.12$)显著高于中等不公平分配者提议的接受率($M = 0.61$, $SD = 0.37$)，$t(30) = -5.93$, $p < 0.001$。对公平分配者提议的接受率($M = 0.95$, $SD = 0.12$)显著高于不公平分配者提议的接受率($M = 0.04$, $SD = 0.18$)，$t(30) = -24.26$, $p < 0.001$。对中等不公平分配者提议的接受率($M = 0.61$, $SD = 0.37$)显著高于不公平分配者提议的接受率($M = 0.04$, $SD = 0.18$)，$t(30) = -8.37$, $p < 0.001$。

不同条件下分配方案所引发的愉悦情绪差异显著，$F(2, 60) = 63.52$，$p < 0.001$, $\eta^2 = 0.68$。公平分配方案所引发的愉悦情绪($M = 6.01$, $SD = 1.47$)显著高于中等不公平分配方案引发的愉悦情绪($M = 4.44$, $SD = 1.51$)，$t(30) = -6.35$, $p < 0.001$。公平分配方案所引发的愉悦情绪($M = 6.01$, $SD = 1.47$)显著高于不公平分配方案所引发的愉悦情绪($M = 3.08$, $SD = 1.40$)，$t(30) = -10.07$, $p < 0.001$。中等不公平分配方案所引发的愉悦情绪($M = 4.44$, $SD = 1.51$)显著高于不公平分配方案所引发的愉悦情绪($M = 3.08$, $SD = 1.40$)，$t(30) = -5.69$, $p < 0.001$。

对不同条件下分配者面孔的公平程度评分差异显著，$F(2, 60) = 24.24$，$p < 0.001$，$\eta^2 = 0.45$，被试对公平分配者的公平程度评分($M = 6.13$, $SD = 2.35$)显著高于对中等不公平分配者($M = 4.29$, $SD = 1.77$)，$t(30) = -3.42$, $p = 0.005$。对公平分配者的公平程度评分($M = 6.13$, $SD = 2.35$)显著高于对不公平分配者的公平程度评分($M = 2.42$, $SD = 1.84$)，$t(30) = -5.82$, $p < 0.001$。对中等不公平分配者的公平程度评分($M = 4.29$, $SD = 1.77$)显著高于对不公平分配者的公平程度评分($M = 2.42$, $SD = 1.84$)，$t(30) = -4.73$, $p < 0.001$。上述结果表明，被试能够将分配者的面孔与不同公平程度之间形成联结。

2.5.2 刺激泛化阶段

混合线性回归分析发现，相对于中等不公平面孔，随着变形面孔与原始不公平面孔知觉相似性的增加，被试更少选择变形面孔进行接下来的信任游戏，不公平面孔联结×面孔相似度 $t(550) = -5.97$, $p < 0.001$。而随着变形面孔与原始公平面孔知觉相似性的增加，被试选择变形面孔比例有所提高，但与中等不公平条件下的差异并未到达显著，$t(550) = 0.64$, $p = 0.524$ (图 4A)。贝叶斯线性回归分析的结果也验证了上述结果，不公平面孔联结×面孔相似度回归系数的 95% *CIs* 为[-0.010, -0.005]不包含 0，公平面孔联结×面孔相似度回归系数的 95% *CIs* 为[-0.001, 0.003]包含 0（图 4B)。此外，为了分析刺激泛化效应是否存在不对称性，将中等不公平条件作为参照水平，对公平和不公平条件下面孔相似性的回归直线斜率进行差异检验。结果发现，不公平条件下的回归直线斜率显著负于公平条件下的回归直线斜率，$z = -6.61$,

$p < 0.001$。上述结果表明，被试对不公平分配者面孔的泛化强度高于对公平分配者面孔的泛化强度。

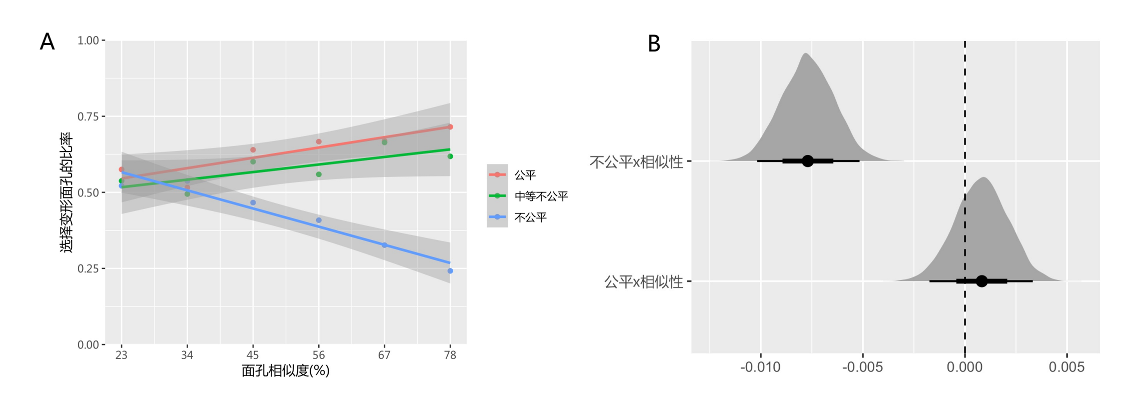


图 4 (A) 不同面孔联结类型下，面孔相似度与选择变形面孔比率的回归分析；(B) 公平/不公平面孔联结×面孔相似度回归系数的后验概率分布及相应的可信度区间 *CIs*。

对信任选择阶段的反应时进行 3 面孔联结类型（公平、中等不公平、不公平）×2 被试选择（信任、不信任）的两因素方差分析[2]。结果表明，面孔联结类型的主效应不显著，$F(2, 48) = 1.48$，$p = 0.237$，$\eta^2 = 0.06$；被试选择的主效应不显著，$F(1, 24) = 3.27$，$p = 0.083$, $\eta^2 = 0.12$。面孔联结类型与被试选择的交互作用显著，$F(2, 48) = 8.91, p < 0.001$，$\eta^2 = 0.27$。简单效应分析表明，在公平和中等不公平条件下，被试选择信任与不信任的反应时差异不显著($F(1, 24) = 1.02, p = 0.322; F(1, 24) = 0.08, p = 0.777$)。在不公平条件下，被试选择信任的反应时($M = 2.64$, $SD = 1.26$)长于选择不信任的反应时($M = 2.04$, $SD = 0.94$)，$F(1, 24) = 20.75$, $p < 0.001$。

[2] 在对反应时数据进行分析时，去除了正负 3 个标准差之外的试次。有 5 名被试在某些条件下仅选择了一个选项，在反应时数据中进行了排除。

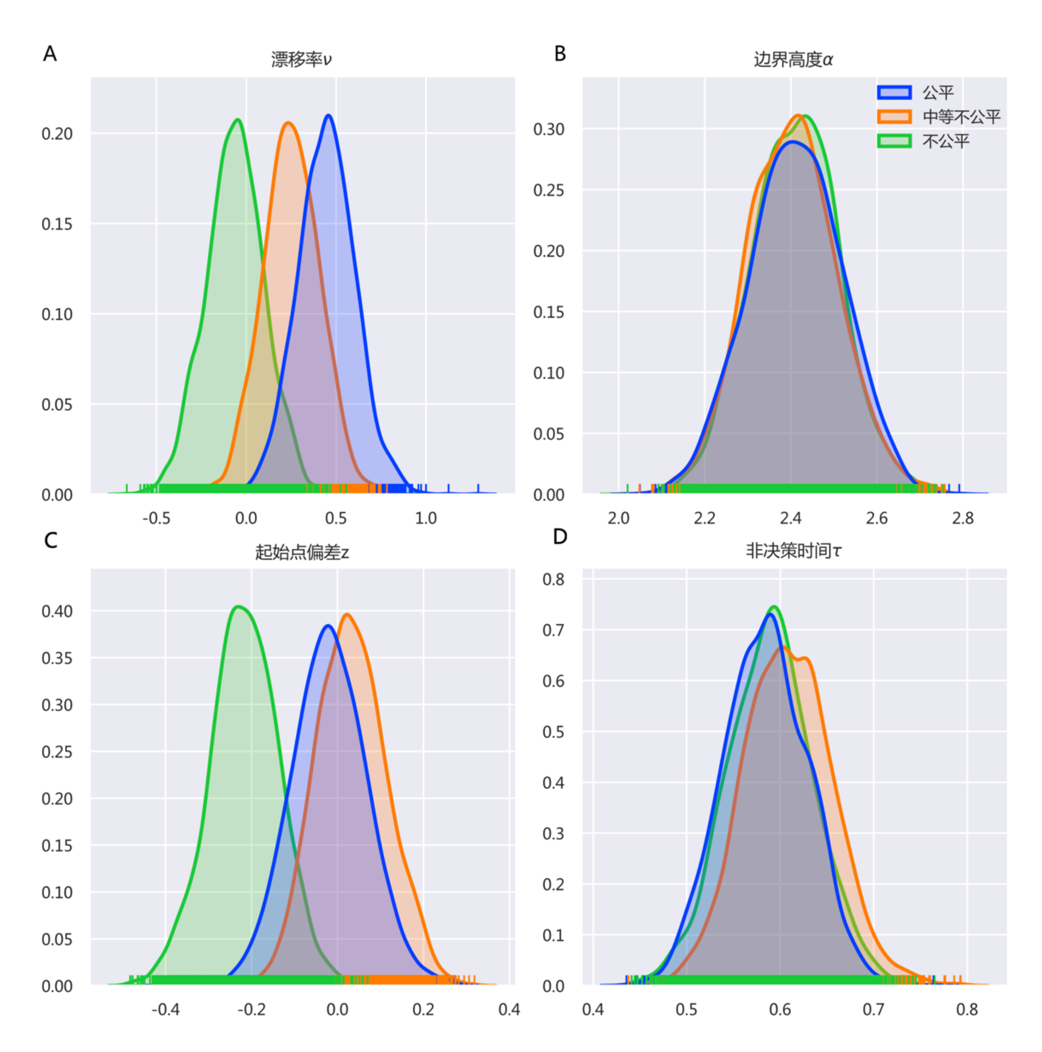


图 5 DDM 的 4 个参数在不同面孔联结类型条件下的概率密度分布。图 A 代表表漂移率v; 图 B 代表边界高度α；图 C 代表边界起始点偏差z；图 D 代表非决策时间τ。

漂移扩散模型分析结果发现，不公平与公平条件之间漂移率v的差异存在较为可靠的证据(M = -0.51, 95% HDI [-0.96, -0.05])，不公平条件下的漂移率v显著小于公平条件下的漂移率v；不公平与中等不公平条件间漂移率v存在差异的证据并不明显(M = -0.31, 95% HDI [-0.75, 0.13]) (见图 5A)。此外，不公平与公平条件间起始点偏差z的差异存在较为可靠的证据(M = -0.06, 95% HDI [-0.12, -0.004])，不公平与中等不公平条件间起始点偏差z的差异也存在较为可靠的证据(M = -0.05, 95% HDI [-0.11, 0.004]) (见图 5C)。三种条件下的边界高度α与非决策时间τ基本重叠，差异分布 95% HDI 均包含 0，出现差异的可能性较小(见图 5B 和图 5D)。

3 实验1b 观察学习下，跨情境的刺激泛化在面孔信任形成中的作用

实验 1a 的结果表明，在直接互动下，被试能够对不同效价的面孔形成联结，进而泛化到不同任务情境下知觉相似的面孔中。实验 1b 在实验 1a 的基础上，考察在观察学习下，是否也能产生跨情境的刺激泛化。

3.1 被试

样本量确定同实验1a。在某高校的学生群体中，招募30名在校大学生，被试平均年龄为20.66($SD = 2.62$)，其中男生7人，女生23人。被试实验前签署知情同意书，实验后给予一定的实验报酬。

3.2 实验材料与程序

实验1b所用实验材料和流程与实验1a基本相同。不同之处仅在于刺激联结阶段，让被试作为第三方观察3名不同公平程度（公平、中等不公平、不公平）的分配者与一个陌生人进行独裁者博弈任务，并在看到分钱方案后对其表现出的公平程度进行评价。在实验1b的刺激联结阶段，3名分配者提出的分配方案同实验1a。刺激联结阶段的具体流程如下：首先，屏幕上会呈现一个“+”，呈现时间为600ms，提示本轮实验任务开始。随后，屏幕上会呈现分配者的面孔图片，被试有2000ms的时间对它进行观察。接着600ms的空屏之后，呈现分配者面孔及其提供给陌生人的分配方案。被试对该分配方案的公平程度进行1~9点(*1 = 非常不公平, 9 = 非常公平*)评定。此外，在30%的试次中，被试还需要对看到该分配方案后的愉悦情绪进行1~9点(*1 = 非常不高兴, 9 = 非常高兴*)评定。在所有试次结束后，呈现3名分配者的面孔，被试需要对每个分配者在先前任务中表现出的公平程度进行1~9点(*1 = 非常不公平, 9 = 非常公平*)评定，以检验被试是否将面孔及其价值之间形成联结。实验1b中刺激泛化阶段的实验流程同实验1a（实验1b具体指导语见补充材料）。

3.3 实验结果

3.3.1 刺激联结阶段

为了检验刺激联结的效果，对被试在不同条件下对分配方案公平程度的评分、分配方案引发的愉悦情绪、以及对分配者面孔的公平程度评分进行单因素被试内方差分析。结果发现，不同条件下对分配方案公平程度的评分存在显著差异，$F(2, 58) = 156.51$，$p < 0.001$, $\eta^2 = 0.84$。对公平分配方案的评分($M = 7.54, SD = 1.15$)显著高于对中等不公平分配方案的评分($M = 4.09, SD = 1.77$，$t(29) = -10.17, p < 0.001$)以及不公平的分配方案的评分($M = 1.83, SD = 1.24$，$t(29) = -16.15, p < 0.001$)。对中等不公平分配方案的评分($M = 4.09, SD = 1.77$)显著高于对不公平分配方案的评分($M = 1.83, SD = 1.24$)，$t(29) = -8.15, p < 0.001$。

不同条件下分配方案所引发的愉悦情绪差异显著，$F(2, 58) = 48.04$，$p < 0.001$, $\eta^2 = 0.62$。公平分配方案所引发的愉悦情绪($M = 6.33, SD = 1.45$)显著高于中等不公平分配方案所引发的愉悦情绪($M = 4.29, SD = 1.24$，$t(29) = -7.28, p < 0.001$)以及不公平分配方案引发的愉悦情绪($M = 3.18, SD = 1.52$，$t(29) = -7.71, p < 0.001$)。中等不公平分配方案所引发的愉悦情绪($M = 4.29, SD = 1.24$)显著高于不公平分配方案引发的愉悦情绪($M = 3.18, SD = 1.52$)，$t(29) = -4.09, p < 0.001$。

对不同条件下分配者面孔的公平程度评分差异显著，$F(2, 58) = 10.47$，$p < 0.001$，$\eta^2 =$

0.27，被试对公平分配者的公平程度评分($M = 6.00$, $SD = 2.57$)高于对中等不公平分配者的公平程度评分($M = 4.93$, $SD = 2.16$)，但差异并未达到显著，$t(29) = -1.55$, $p = 0.398$。对公平分配者的公平程度评分($M = 6.00$, $SD = 2.57$)显著高于对不公平分配者的公平程度评分($M = 3.27$, $SD = 1.98$)，$t(29) = -4.69$, $p < 0.001$。对中等不公平分配者的公平程度评分($M = 4.93$, $SD = 2.16$)显著高于对不公平分配者的公平程度评分($M = 3.27$, $SD = 1.98$)，$t(29) = -3.20$, $p = 0.003$。上述结果表明，被试能够将分配者的面孔与不同公平程度之间形成联结。

3.3.2 刺激泛化阶段

混合线性回归分析发现，相对于中等不公平面孔，随着与不公平面孔知觉相似性的增加，被试更少选择变形面孔进行接下来的信任游戏，不公平面孔联结×面孔相似度 $t(532) = -3.81$, $p < 0.001$。此外，相对于中等不公平，随着与原始公平面孔知觉相似性的增加，被试更多选择变形面孔进行接下来的信任游戏，$t(532) = 2.12$, $p = 0.004$ (图 6A)。贝叶斯线性回归分析的结果也验证了上述结果，不公平面孔联结×面孔相似度回归系数的 95% *CIs* 为[-0.008, -0.003]不包含 0，公平面孔联结×面孔相似度回归系数的 95% *CIs* 为[0.0002, 0.006]也不包含 0（图 6B)。与实验 1a 相同，将中等不公平条件作为参照水平，对公平和不公平条件下面孔相似性的回归直线斜率进行差异检验。结果发现，不公平面孔的回归直线斜率显著负于公平面孔的回归直线斜率，$z = -5.93$, $p < 0.001$。上述结果表明，在观察学习下，被试对不公平分配者面孔的泛化强度仍高于对公平分配者面孔的泛化强度。

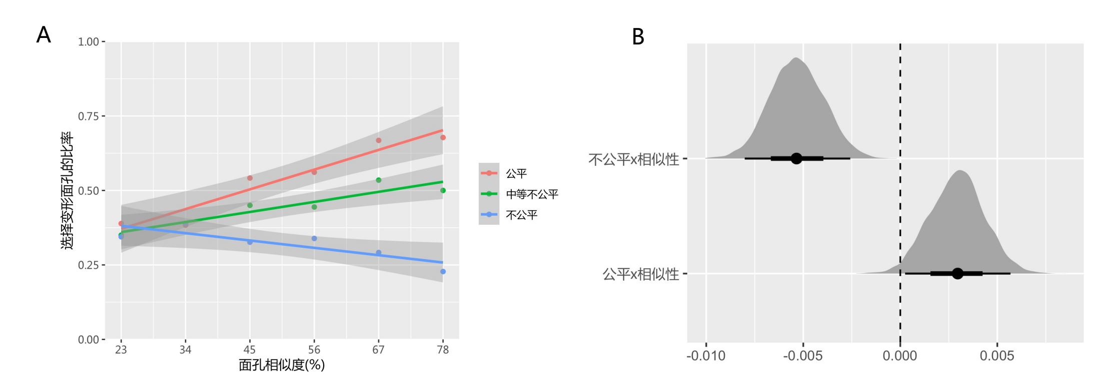


图 6 (A) 同面孔联结类型下，面孔相似度与选择变形面孔比率的回归分析；(B) 公平/不公平面孔联结×面孔相似度回归系数的后验概率分布及相应的可信度区间 *CIs*。

对信任选择阶段的反应时进行 3 面孔联结类型（公平、中等不公平、不公平）×2 被试选择（信任、不信任）的两因素方差分析[3]。结果表明，面孔联结类型的主效应不显著，$F(2, 50) = 1.40$，$p = 0.257$，$\eta^2 = 0.05$；被试选择的主效应不显著，$F(1, 25) = 1.99$, $p = 0.171$，$\eta^2 = 0.07$。面孔联结类型与被试选择的交互作用不显著，$F(2, 50) = 1.25$, $p = 0.294$，$\eta^2 = 0.05$。漂移扩散模型分析结果发现，不公平与公平条件间漂移率v的差异存在较为可靠的证据(M = -

[3] 在对反应时数据进行分析时，去除了正负 3 个标准差之外的试次。有 4 名被试在某些条件下仅选择了一个选项，在反应时数据中进行了排除。

0.60, 95% HDI [-0.97, -0.24])，不公平条件下的漂移率v显著小于公平条件下的漂移率v；不公平与中等不公平条件间的漂移率v的差异也存在较为可靠的证据(M = -0.37, 95% HDI [-0.75, -0.03])，不公平条件下的漂移率v显著小于中等不公平条件下的漂移率v(见图 7A)。而三种条件下的边界高度α、始点偏差z与非决策时间τ基本重叠，不同条件间差异分布的 95% HDI 均包含 0，出现差异的可能性较小(见图 7B、图 7C 和图 7D)。

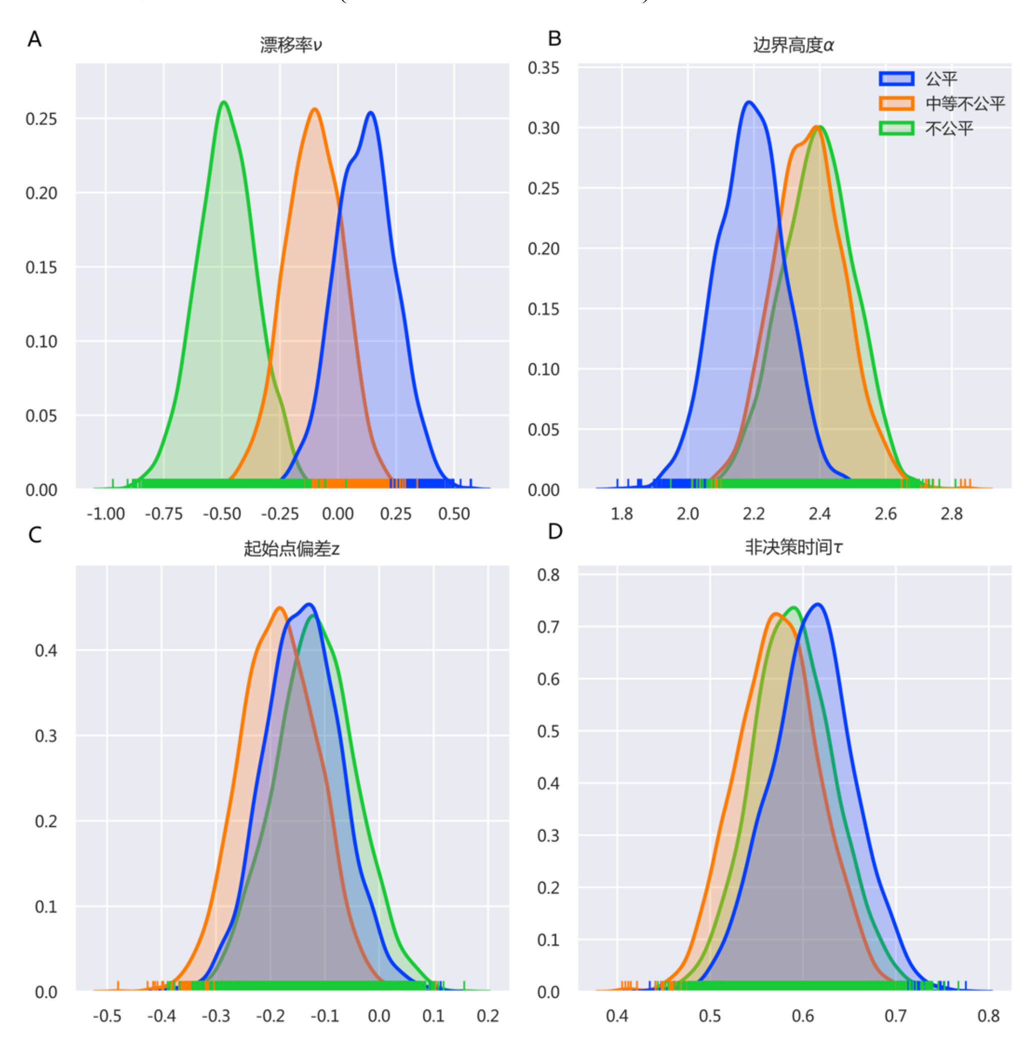


图 7 DDM 的 4 个参数在不同面孔联结类型条件下的概率密度分布。图 A 代表表漂移率v; 图 B 代表边界高度α；图 C 代表边界起始点偏差z；图 D 代表非决策时间τ。

4 实验2 行为意图在跨情境的刺激泛化中的作用

实验 1a 和实验 1b 的结果表明，无论在直接互动还是在观察学习下，被试均能对不同效价面孔形成联结，并泛化到不同任务情境下知觉相似的面孔中，但上述两个实验仅证明了在有行为意图的条件产生了刺激泛化。实验 2 旨在考察在无意图条件下，个体是否也会产生上述刺激泛化效应。

4.1 被试

样本量确定同实验 1。在某高校招募 30 名在校大学生，被试平均年龄为 19.86(SD = 2.43)，其中男生 11 人，女生 19 人。被试实验前签署知情同意书，实验后给予一定的实验报酬。

4.2 实验材料与任务

实验 2 所用实验材料和流程与实验 1a 基本相同。不同之处仅在于刺激联结阶段，由计算机提出具有不同公平程度的分配方案（公平、中等不公平、不公平）给被试和搭档，被试决定是否接受该分配方案。如果接受，则按计算机提出的分配方案进行分配；如果不接受，则被试和陌生人均获得 0 元。通过上述操作，无论分配结果公平与否，都与交往对象的行为意图无关；但交往对象的面孔依然与不同效价的分配结果匹配呈现，只是这些结果并非由交往对象的意图导致。因此，上述操纵可以排除互动对象的意图，同时不引入其他可能的无关变量，考察意图感知在刺激泛化中的作用。

刺激联结阶段的具体流程如下：首先，屏幕上会呈现一个“+”，呈现时间为 600ms，提示本轮实验任务开始。随后，屏幕上会呈现计算机的轮廓图片，被试有 2000ms 的时间对它进行观察。接着 600ms 的空屏之后，呈现陌生人面孔及计算机提供的分配方案。被试需要决定是否接受该分配方案，如果接受按“*F*”键，拒绝按“*J*”键。被试做出按键反应后，会对被试的选择结果进行反馈，反馈呈现 2000ms。此外，在 30%的试次中，被试还需要评定看到该分配方案后的愉悦情绪(*1 = 非常不高兴, 9 = 非常高兴*)。最后，呈现给被试陌生人的面孔，要求被试评定与该人进行分钱游戏时，计算机生成的分配方案的公平程度(*1 = 非常不公平, 9 = 非常公平*)，以检验被试是否将面孔及其价值之间形成联结。实验 2 中刺激泛化阶段的实验流程同实验 1a(实验 2 具体指导语见补充材料)。

4.3 实验结果

4.3.1 刺激联结阶段

对被试在不同条件下对计算机提议的接受率、分配方案引发的愉悦情绪、以及与不同面孔进行分配时计算机提出分配方案的公平程度评分进行单因素被试内方差分析。结果发现，对不同条件下计算机提议的接受率差异显著，$F(2, 58) = 143.69$，$p < 0.001$, $\eta^2 = 0.83$。对公平分配方案的接受率($M = 0.95$, $SD = 0.09$)显著高于对中等不公平分配方案的接受率($M = 0.62$, $SD = 0.34$，$t(29) = -5.99$, $p < 0.001$)以及不公平分配方案的接受率($M = 0.09$, $SD = 0.21$，$t(29) = -22.15$, $p < 0.001$)。对中等不公平分配方案的接受率($M = 0.62$, $SD = 0.34$)显著高于对不公平分配方案的接受率($M = 0.09$, $SD = 0.21$)，$t(29) = -9.21$, $p < 0.001$。

不同条件下分配方案所引发的愉悦情绪差异显著，$F(2, 58) = 34.37$，$p < 0.001$, $\eta^2 = 0.54$。公平分配方案所引发的愉悦情绪($M = 5.30$, $SD = 1.23$)显著高于中等不公平分配方案($M = 4.00$, $SD = 1.38$，$t(29) = -5.00$, $p < 0.001$)以及不公平分配方案引发的愉悦情绪($M = 3.33$, $SD = 1.18$，$t(29) = -8.57$, $p < 0.001$)。中等不公平分配方案所引发的愉悦情绪($M = 4.00$, $SD = 1.38$)显著高于不公平分配方案引发的愉悦情绪($M = 3.33$, $SD = 1.18$)，$t(29) = -2.86$, $p = 0.023$。然而，与不同分配者面孔进行互动时，对计算机生成的分配方案公平程度评分差异不显著，$F(2, 58) = 1.36$，$p = 0.266$，$\eta^2 = 0.05$。上述结果表明，在无意图条件下，被试虽然更多地拒绝了计算机生成的不公平分配方案，但并没有将互动对象的面孔与不同公平程度之间形成价值联结。

4.3.2 刺激泛化阶段

混合线性回归分析发现，相对于中等不公平面孔，随着与原始不公平面孔知觉相似性的增加，被试选择变形面孔进行信任游戏的比率没有显著变化，$t(532) = -0.12, p = 0.902$。此外，相对于中等不公平面孔，随着与原始公平面孔知觉相似性的增加，被试选择变形面孔进行信任游戏的比率也没有显著变化，$t(532) = 1.78, p = 0.075$ (图 8A) (我们也综合分析了实验 1a 与实验 2 的数据，结果见补充材料)。贝叶斯线性回归分析的结果也验证了上述结果，不公平面孔联结×面孔相似度回归系数的 95% *CIs* 为[-0.003, 0.002]包含 0，公平面孔联结×面孔相似度回归系数的 95% *CIs* 为[-0.001, 0.005]包含 0 (图 8B)。

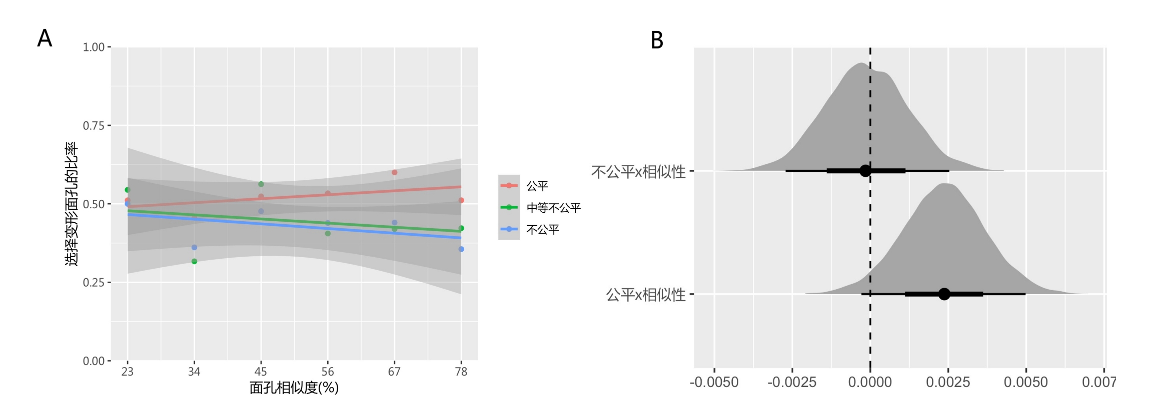


图 8 (A) 不同面孔联结类型下，面孔相似度与选择变形面孔比率的回归分析；(B) 公平/不公平面孔联结×面孔相似度回归系数的后验概率分布及相应的可信度区间 *CIs*。

对信任选择阶段的反应时进行 3 面孔联结类型（公平、中等不公平、不公平）×2 被试选择（信任、不信任）的两因素方差分析[4]。结果表明，面孔联结类型的主效应不显著，$F(2, 54) = 1.30$，$p = 0.280$，$\eta^2 = 0.05$；被试选择的主效应不显著，$F(1, 27) = 1.91, p = 0.178$，$\eta^2 = 0.01$。面孔联结类型与被试选择的交互作用不显著，$F(2, 54) = 0.17, p = 0.848$，$\eta^2 = 0.01$。采用同实验 1a 的方法对反应时数据进行了 DDM 分析。结果发现，三种条件下的漂移率v、边界高度α、始点偏差z与非决策时间τ基本重叠，不同条件间差异分布的 95% HDI 均包含 0，出现差异的可能性较小(见图 9)。

[4] 在对反应时数据进行分析时，去除了正负 3 个标准差之外的试次。有 2 名被试在某些条件下仅选择了一个选项，在反应时数据中进行了排除。

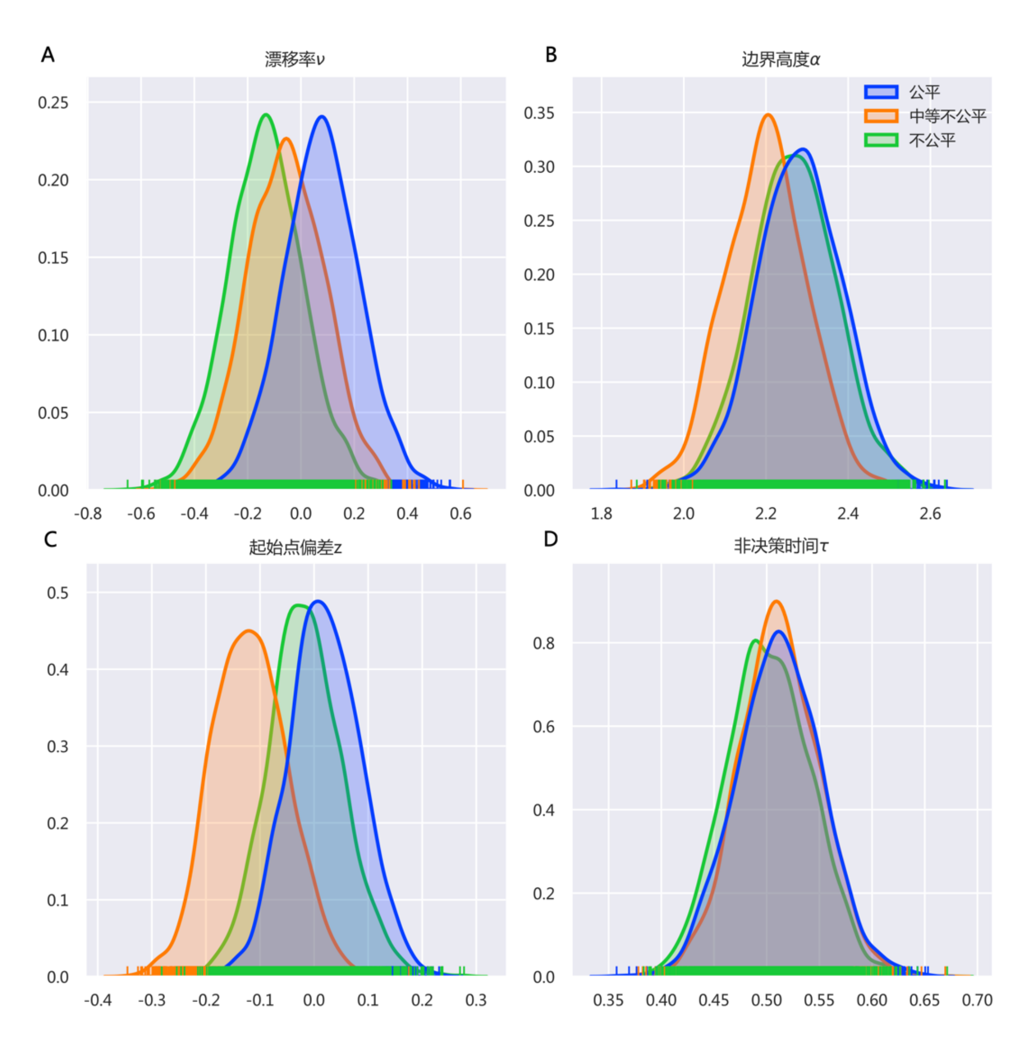


图 9 DDM 的 4 个参数在不同面孔联结类型条件下的概率密度分布。图 A 代表表漂移率ν; 图 B 代表边界高度α；图 C 代表边界起始点偏差z；图 D 代表非决策时间τ。

5 讨论

信任在社会的各个层面都很重要，常常被认为是维系社会关系的“胶水”(Sullivan & Transue, 1999; Zak & Knack, 2001)。然而，在缺乏直接经验的情况下，人们如何选择信任陌生人的认知机制目前还知之甚少。泛化效应是对陌生面孔印象形成的重要途径，先前的研究发现泛化效应存在于同种属性之间或相同任务情境下。例如，Kocsor 和 Bereczkei (2017)研究发现，当学习面孔与积极特质配对时，相似的陌生面孔会获得更高的积极评价；当学习面孔与消极特质配对时，相似的陌生面孔获得的评价也更消极。Zebrowitz 和 Montepare (2008)发现，对熟悉面孔的喜爱度也会发生泛化进而影响对相似的陌生面孔的喜爱度。Feldmanhall 等人(2018)发现个体会更信任那些与在先前互动中值得信赖的面孔相似的面孔，不信任与在先前互动中不值得信赖的面孔相似的面孔。

在以往研究的基础上，本研究进一步发现，刺激泛化效应在跨情境之间也会发生，联结阶段互动对象的（不）公平特质会被泛化到随后的信任决策中。这表明，在缺乏直接经验或声誉信息的情况下，个体对陌生人进行信任决策时，会采用联结学习机制将不同情境中习得的声誉信息泛化到新的互动情境中，进而指导其做出信任决策。由于刺激很少以完全相同的

形式出现，基于相似性的跨情境泛化机制具有很强的适应性。在日常生活中，我们常常遇到一些陌生人，对他们可信度的判断往往缺乏相应的信息。在这种缺乏直接经验的情况下，我们做出的决定依赖于从过去经验中进行归纳的能力。这种联结学习机制可以利用先前互动中的学习，减少新情境中与陌生人互动的不确定性，从而促进潜在的适应性信任决策(FeldmanHall et al., 2018)。

除了直接互动情境，在观察学习情境下，我们也发现了面孔信任形成中跨情境的刺激泛化效应。这表明，个体通过观察他人的行动结果，能够间接的习得刺激的价值，进而采用刺激泛化的机制指导未来的行为。通过观察他人在社会交互中的行为表现，人们能够形成对他人的声誉表征，并在随后的互动中根据他人声誉调整相应的行为(Milinski, 2016)。间接的观察学习在自然界中广泛存在, 对个体适应复杂的社会环境以及优化社会决策有着重要的意义。例如，恒河猴在观察到同类对蛇的恐惧反应后，能迅速习得同样的对蛇的恐惧反应(Mineka & Ohman, 2002)。人类社会中也普遍存在通过观察学习的方式习得恐惧反应的现象(Olsson et al., 2007)。这种学习方式，不仅帮助观察者避免亲自执行行为可能带来的消极后果，也让观察者有效地习得使自身获益最大化的行为(Frith & Frith, 2012)。面孔信任形成中跨情境的刺激泛化效应表明，个体在进行信任决策时，采用联结学习机制将不同情境中习得的刺激价值联结泛化到新的互动情境中，进而指导其随后的信任决策。这种基于相似性的泛化机制具有很强的适应性，能使个体从最小的学习中获得更高价值。当然，需要指出是，基于相似性的泛化机制也存在一定的局限，人们会将某种特定情境中的刺激价值联结泛化到其他相关甚至无关的情境中。这可能导致个体忽视当下的情境，对他人产生错误的判断。过分泛化还可能导致刻板印象和偏见的形成，具有潜在的消极作用(Allidina & Cunningham, 2021)。

通过基于决策选择和反应时的漂移扩散模型分析，我们进一步探究了跨情境间的刺激泛化在面孔信任形成中的认知计算机制。结果发现，相对于公平和中等不公平条件，在不公平条件下，个体选择信任的反应时长于选择不信任的反应时。这表明，在不公平条件下，个体做出信任选择更加困难，需要更多的认知加工时间。漂移扩散模型分析发现，不公平与中等不公平或公平条件之间差异主要体现在漂移率v，不公平条件下的漂移率v显著小于中等不公平或公平条件下的漂移率v。DDM 中的漂移率v代表了累积某一选择证据的速率，个体倾向于某一选项的偏好越强烈，信息向该选项积累的速度就越快(Forstmann et al., 2016)。上述表明，在对与不公平分配者面孔相似的陌生面孔进行信任决策时，个体更倾向累积其不值得信任的证据，进而更快地做出不信任的决策。虽然，以往 DDM 主要用于感知决策的研究，但最近它在社会决策领域获得了很多关注，成功地解释了群体对个体利他惩罚决策的影响(Son et al., 2019)，以及利他决策和道德伤害决策的认知机制(Germar et al., 2014; Yu et al., 2021)。我们首次采用 DDM 探究了跨情境下的刺激泛化影响信任决策的认知过程，展示了其在探讨社会决策中的应用价值。

本研究还发现，面孔信任形成中跨情境的刺激泛化效应具有不对称性，相比于公平面孔，

个体对不公平面孔产生了更强的刺激泛化；当陌生面孔与学习阶段不公平面孔具有较小的相似度时，个体就倾向于认为陌生面孔是不值得信任的。这表明，个体更加注重先前社会互动中的负性道德信息，用来指导其进行随后的适应性信任决策。Grady 等人(2007)研究发现，相比于正性价值的面孔，个体更倾向于优先关注具有负面价值的面孔。在对社会行为的观察学习中也普遍存在消极偏向(Baumeister et al., 2001; Rozin & Royzman, 2001)，这可能是由于忽视环境中的潜在危险所造成的危害可能远远大于错失一个机会，因此环境中的消极信息比积极或中性信息更容易获得注意。

最后，本研究还发现了行为意图在形成刺激泛化效应中的重要作用。Ma 等人(2015)从公平感的角度引入有意图的最后通牒范式来厘清不同意图对互惠行为的影响，结果发现，在提供公平或不公平的分配方案条件下，被试对不公平提议的拒绝率受到提议者意图的调节。意图对行为结果的调节作用，在道德判断(Cushman et al., 2013)、直接互惠(Vaish et al., 2018)、广义互惠(Sun et al., 2020)等其他领域也得到了验证。本研究首次将行为意图的影响拓展到刺激泛化领域，表明了仅在有意图的条件下，个体才将结果效价与面孔刺激之间形成联结，进而采用刺激泛化的机制指导其随后的信任决策。

采用联结学习机制从过去的经验中捕捉互动对象的道德信息，进而指导个体随后的信任决策。这些研究发现有助于理解个体对陌生人信任的内在心理机制，对丰富学习泛化领域研究具有重要意义。本研究也存在着一些不足，首先，本研究并未对个体差异变量（如公平敏感性、社会价值取向）进行考察，这些变量可能也在泛化效应中起着调节作用。其次，本研究的任务与典型刺激泛化任务不同，在典型刺激泛化任务中，新刺激通常被认为与原始刺激有明显重叠。而在本研究中，虽然刺激泛化阶段的变形面孔与刺激联结阶段的学习面孔具有不同的相似性，不过被试事后主观报告，他们在实验过程中并未意识到变形面孔与学习面孔之间的关联，这在一定程度上反映了面孔相似性操作的内隐性。在这种内隐情形下，个体对变形面孔的信任仍然受到过去经验的影响，表明面孔信任中的刺激泛化具有一定的内隐性(罗秋铃 等, 2020; Verosky & Todorov, 2013)。然而这种内隐学习泛化的认知机制和神经基础尚不清楚，未来研究可借助认知神经技术进行探索。

6 结论

基于联结学习理论，本研究考察了跨情境的刺激泛化在面孔信任形成中的作用。研究结果表明，随着被信任者的面孔与先前互动中公平（不公平）分配者面孔相似度的增加，个体对其信任程度逐渐增加（降低）；并且对不公平分配者面孔的泛化强度高于对公平面孔的泛化强度。此外，行为意图在刺激泛化效应的产生中起到调节作用；在无意图条件下，上述跨情境的刺激泛化效应消失。

参考文献

Allidina, S., & Cunningham, W. A. (2021). Avoidance begets avoidance: a computational account of negative stereotype persistence. *Journal of Experimental Psychology: General*, *150*(10), 2078–2099.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412.

Bateson, M., Brilot, B., & Nettle, D. (2011). Anxiety: an evolutionary approach. *The Canadian Journal of Psychiatry*, *56*(12), 707–715.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*(4), 323–370.

Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2011). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*, 1–28.

Buckholtz, J. W., Asplund, C. L., Dux, P. E., Zald, D. H., Gore, J. C., Jones, O. D., & Marois, R. (2008). The neural correlates of third-party punishment. *Neuron*, *60*(5), 930–940.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*, 1–28.

Chang, L. J., & Sanfey, A. G. (2013). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience*, *8*(3), 277–284.

Cushman, F., Dreber, A., Wang, Y., & Costa, J. (2009). Accidental outcomes guide punishment in a “trembling hand” game. *PloS one*, *4*(8), e6699.

Cushman, F., Sheketoff, R., Wharton, S., & Carey, S. (2013). The development of intent-based moral judgment. *Cognition*, *127*(1), 6–21.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473–492.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*(2), 268–298.

Earley, R. L. (2010). Social eavesdropping and the evolution of conditional cooperation and cheating strategies. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *365*(1553), 2675–2686.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review*, *88*(4), 848–881.

Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, *6*, 148.

FeldmanHall, O., Dunsmoor, J. E., Tompary, A., Hunter, L. E., Todorov, A., & Phelps, E. A. (2018). Stimulus generalization as a mechanism for learning to trust. *Proceedings of the National Academy of Sciences*, *115*(7), 1690–1697.

Forstmann, B. U., Ratcliff, R., & Wagenmakers, E.-J. (2016). Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions. *Annual Review of Psychology*, *67*(1), 641–666.

Frith, C. D., & Frith, U. (2012). Mechanisms of social cognition. *Annual Review of Psychology*, *63*, 287–313.

Gao, Q. L., & Zhou, Y. (2021). Psychological and neural mechanisms of trust formation: A perspective from computational modeling based on the decision of investor in the trust game. *Advances in Psychological Science*, *29*(1), 178–189.

[高青林, 周媛. (2021). 计算模型视角下信任形成的心理和神经机制——基于信任博弈中投资者的角度. *心理科学进展*, *29*(1), 178–189.]

Gao, W., Cao, B., Shan, S., Chen, X., Zhou, D., Zhang, X., & Zhao, D. (2007). The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *38*(1), 149–161.

Gawronski, B., & Quinn, K. A. (2013). Guilty by mere similarity: assimilative effects of facial resemblance on automatic evaluation. *Journal of Experimental Social Psychology*, *49*(1), 120–125.

Germar, M., Schlemmer, A., Krug, K., Voss, A., & Mojzisch, A. (2014). Social influence and perceptual decision making: A diffusion model analysis. *Personality and Social Psychology Bulletin*, *40*(2), 217–231.

Grady, C. L., Hongwanishkul, D., Keightley, M., Lee, W., & Hasher, L. (2007). The effect of age on memory for emotional faces. *Neuropsychology*, *21*(3), 371–380.

Johnson, D. J., Hopwood, C. J., Cesario, J., & Pleskac, T. J. (2017). Advancing research on cognitive processes in social and personality psychology: A hierarchical drift diffusion model primer. *Social Psychological and Personality Science*, *8*(4), 413–423.

Kocsor, F., & Bereczkei, T. (2017). First impressions of strangers rely on generalization of behavioral traits associated with previously seen facial features. *Current Psychology*, *36*(3), 385–391.

Kraus, M. W., Chen, S., Lee, V. A., & Straus, L. D. (2010). Transference occurs across group boundaries. *Journal of Experimental Social Psychology*, *46*(6), 1067–1073.

Luo, Q., Huang, L., Hou, Q., Aimaitijiang, R., Zhou, M., Zhou, X., & Chen, S. (2020). Generalization effect of gossip on interpersonal trust. *Journal of Psychological Science*, *43*(1), 165–171.

[罗秋铃, 黄露霖, 侯庆辉, 热米拉·艾买提江, 周梦哲, 周晓林, 陈双 (2020). 传言对人际信任影响的泛化效应. *心理科学*, *43*(1), 165–171.]

Ma, Q., Meng, L., Zhang, Z., Xu, Q., Wang, Y., & Shen, Q. (2015). You did not mean it: perceived good intentions alleviate sense of unfairness. *International Journal of Psychophysiology*, *96*(3), 183–190.

Mcauliffe, K., Jordan, J. J., & Warneken, F. (2015). Costly third-party punishment in young children. *Cognition*, *134*, 1–10.

Milinski, M. (2016). Reputation, a universal currency for human social interactions. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1687).

Mineka, S., & Ohman, A. (2002). Phobias and preparedness: the selective, automatic, and encapsulated nature of fear. *Biological Psychiatry*, *52*(10), 927–937.

Moretti, L., & Di Pellegrino, G. (2010). Disgust selectively modulates reciprocal fairness in economic interactions. *Emotion*, *10*(2), 169–180.

Offerman, T. (2002). Hurting hurts more than helping helps. *European Economic Review*, *46*(8), 1423–1437.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: the neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*, *2*(1), 3–11.

Olsson, A., Knapska, E., & Lindström, B. (2020). The neural and computational systems of social learning. *Nature Reviews Neuroscience*, *21*(4), 197–212.

Peirce, J. W. (2009). Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics*, *2*, 10.

Radell, M. L., Sanchez, R., Weinflash, N., & Myers, C. E. (2016). The personality trait of behavioral inhibition modulates perceptions of moral character and performance during the trust game: behavioral results and computational modeling. *PeerJ*, *4*, e1631.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Rescorla, R. A., & Solomon, R. L. (1967). Two-process learning theory: relationships between pavlovian conditioning and instrumental learning. *Psychological Review*, *74*(3), 151–182.

Rescorla, R. A. (1976). Stimulus generalization: some predictions from a model of Pavlovian conditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, *2*(1), 88–96.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., & Camerer, C. (1998). Not so different after all: A cross-discipline view of trust. *Academy of Management Review*, *23*(3), 393–404.

Rozin, P., & Royzman, E. B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review*, *5*(4), 296–320.

Ruff, C. C., & Fehr, E. (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience*, *15*(8), 549–562.

Schechtman, E., Laufer, O., & Paz, R. (2010). Negative valence widens generalization of learning. *Journal of Neuroscience*, *30*(31), 10460–10464.

Singer, T., Seymour, B., O'Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, *439*(7), 466–469.

Shang, J., & Li, Y. (2020). The effects of participants' sex and the facial trustworthiness of proposers on third-party decision-making in a dictator game. *PsyCh Journal*, *9*(6), 877–884.

Son, J.-Y., Bhandari, A., & FeldmanHall, O. (2019). Crowdsourcing punishment: Individuals reference group preferences to inform their own punitive decisions. *Scientific Reports*, *9*(1), 11625.

Sullivan, J. L., & Transue, J. E. (1999). The psychological underpinnings of democracy: A selective review of research on political tolerance, interpersonal trust, and social capital. *Annual Review of Psychology*, *50*(1), 625–650.

Sun, Z., Ye, C., He, Z., & Yu, W. (2020). Behavioral intention promotes generalized reciprocity: evidence from the dictator game. *Frontiers in Psychology*, *11*, 772.

Sutter, M. (2007). Outcomes versus intentions: on the nature of fair behavior and its development with age. *Journal of Economic Psychology*, *28*(1), 69–78.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social Cognitive and Affective Neuroscience*, *3*(2), 119–127.

Vaish, A., Hepach, R., & Tomasello, M. (2018). The specificity of reciprocity: young children reciprocate more generously to those who intentionally benefit them. *Journal of Experimental Child Psychology*, *167*, 336–353.

Van't Wout, M., Kahn, R. S., Sanfey, A. G., & Aleman, A. (2006). Affective state and decision-making in the ultimatum game. *Experimental Brain Research*, *169*(4), 564–568.

Verosky, S. C., & Todorov, A. (2010). Generalization of affective learning about faces to perceptually similar faces. *Psychological Science*, *21*(6), 779–785.

Verosky, S. C., & Todorov, A. (2013). When physical similarity matters: Mechanisms underlying affective learning generalization to the evaluation of novel faces. *Journal of Experimental Social Psychology*, *49*(4), 661–669.

Wang, T., Chen, Y., & Lu, J. (2020). The generalization effect in gap evaluation: How large is the gap between you and me? *Acta Psychologica Sinica*, *52*(11), 1327–1339.

[王天鸿, 陈宇琦, 陆静怡.(2020). 差距知觉的泛化效应：我和你之间的差距有多大？*心理学报*, *52*(11), 1327–1339.]

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*.

Willis, J., & Todorov, A. (2006). First impressions: Making up your mind after a 100-ms exposure to a face. *Psychological Science*, *17*(7), 592–598.

Wilson, R. K., & Eckel, C. C. (2006). Judging a book by its cover: Beauty and expectations in the trust game. *Political Research Quarterly*, *59*(2), 189–202.

Xiang, T., Lohrenz, T., & Montague, P. R. (2013). Computational substrates of norms and their violations during social exchange. *Journal of Neuroscience*, *33*(3), 1099–1108.

Yu, H., Siegel, J. Z., Clithero, J. A., & Crockett, M. J. (2021). How peer influence shapes value computation in moral decision-making. *Cognition*, *211*, 104641.

Zak, P. J., & Knack, S. (2001). Trust and growth. *The Economic Journal*, *111*(470), 295–321.

Zebrowitz, L. A., & Montepare, J. M. (2008). Social psychological face perception: Why appearance matters. *Social and Personality Psychology Compass*, *2*(3), 1497–1517.

Zhang, Y., Li, H., & Wu, Y. (2020). The application of computational modelling in the studies of moral cognition. *Advances in Psychological Science*, *28*(7), 1042–1055.

[张银花, 李红, 吴寅. (2020). 计算模型在道德认知研究中的应用. *心理科学进展*, *28*(7), 1042–1055.]

Zheng, X., Guo, W., Chen, M., Jin, J., & Yin, J. (2020). Influence of the valence of social actions on attentional capture: Focus on helping and hindering actions. Acta Psychologica Sinica, 52(5), 584–596.

[郑旭涛, 郭文姣, 陈满, 金佳, 尹军. (2020). 社会行为的效价信息对注意捕获的影响: 基于帮助和阻碍行为的探讨. *心理学报*, *52*(5), 584–596.]

The role of cross-situational stimulus generalization in the formation of trust towards face: a perspective based on direct and observational learning

Abstract

YUAN Bo, WANG Xiaoping, YIN Jun, LI Weiqiang

(*Department of Psychology, Ningbo University, Ningbo, 315211, China*)

How do humans learn to trust unfamiliar others? Decisions in the absence of direct knowledge rely on our ability to generalize from past experiences and are often shaped by the degree of similarity between prior experience and novel situations. A previous study suggested that people prefer to trust toward strangers who resemble the original player they previously learned was trustworthy and avoid trusting toward strangers resembling the untrustworthy player. However, it is still unclear whether this stimulus generalization effect exists across different situations, and the role of intention perception in this effect. Here, we leverage a stimulus generalization framework to examine how perceptual similarity between known individuals and unfamiliar strangers across

different interactive situations shapes people's trust toward strangers. Given that the strong adaptability of the stimulus generalization mechanism, we assume that the faces associated with different degrees of unfairness will affect the individual's trust towards similar unfamiliar faces, and intention perception modulates this process.

Three experiments were conducted to examine the above hypothesis. In Experiment 1a and Experiment 1b, participants play or observe an iterative ultimatum game with three partners who exhibit highly unfair, medium unfair, or highly fair behavior. After learning who was the fair/unfair allocator, participants select new partners for a trust game. Unbeknownst to participants, each potential new partner was parametrically morphed with one of the three original players. In Experiment 2, participants play a similar iterative ultimatum game with three partners, nevertheless the allocations were generated by a computer algorithm which excludes the intention of the allocator.

A mixed linear regression was conducted, with both (un)fairness type (whether faces were morphed with the original fair, medium unfair, unfair allocator' face) and perceptual similarity (increasing similarity to the original face, 23%, 34%, 45%, 56%, 67%, 78%) were entered as predictors of choosing to play with the morphed face. The result of Experiment 1a and Experiment 1b show that compared with the medium unfair condition, as the perceptual similarity between the morphed trustee's face and the face of the fair (unfair) allocator in the previous interaction increases, the degree of trust (distrust) towards the trustee gradually increases. In addition, this effect is asymmetrical, participants preferentially avoided more the unfair morphs in comparison with the fair morphs. This suggests an asymmetric overgeneralization toward individuals perceived to be morally aversive. Using Drift-Diffusion Modeling (DDM), we found that the drift rate v under unfair condition is significantly smaller than that under medium unfair or fair conditions, and most of them are in the range of less than 0. This suggests that individuals are more likely to accumulate evidence of distrust when making trust decisions about unfamiliar faces that are similar to the allocator who was unfair in previous interactions. In Experiment 2, under an unintentional situation, the above-mentioned cross-situational generalization effect disappeared.

Together, our results demonstrate that the individuals use the associative learning mechanism to capture the moral information of the interactive objects from the past experience, and then guides subsequent trust decision-making. This mechanism draws on prior learning to reduce the uncertainty associated with strangers, ultimately facilitating potentially adaptive decisions to trust, or withhold trust from unfamiliar others.

Key words trust formation, associative learning, stimulus generalization, behavioral intention, Drift-Diffusion Modeling